5. Objectives

- approximation
- maximum likelihood estimation
- nonparametric distribution estimation
- discrimination

Residuals and approximation

consider a set of m (possibly nonlinear) equations in n variables $x \in \mathbf{R}^n$:

$$f_i(x) = b_i, \quad i = 1, \ldots, m$$

- $f_i \colon \mathbf{R}^n \to \mathbf{R}$, $b \in \mathbf{R}^m$ are given
- sometimes written as $f(x) = b$, where $f(x) = (f_1(x), \ldots, f_m(x))$

our goal is to

- find x such that $f(x) = b$, *i.e.*, solve the system, or
- find x such that $f(x) \approx b$ in the sense that the *residual*

$$r = f(x) - b$$

  is small under some metric

examples:

- when $f(x) = Ax$ for some $A \in \mathbf{R}^{m\times n}$, we have the *linear model*

$$Ax = b$$

  - residual $r = Ax - b$
  - solvable if $b \in \mathcal{R}(A)$, *i.e.*, can be expressed as a linear combination of the columns of A

- *matrix factorization* problem

$$XY = B$$

  - $f(X, Y) = XY$, $X \in \mathbf{R}^{m\times k}$, $Y \in \mathbf{R}^{k\times n}$, $B \in \mathbf{R}^{m\times n}$
  - residual $R = XY - B$
  - solvable if $k \geq \mathbf{rank}\, B$
  - reduces to linear model if either X or Y is fixed

Norm approximation

$$\text{minimize} \quad \|r\| = \|f(x) - b\|$$

- variable: $x \in \mathbf{R}^n$
- $\|\cdot\|$ is a norm on $\mathbf{R}^m$
- optimal value $\|r^\star\| = 0$ if $f(x) = b$ is solvable
- $x^\star$ called an *approximate solution* of $f(x) = b$ if $\|r^\star\| > 0$

interpretation when $f(x) = Ax$ ($A \in \mathbf{R}^{m\times n}$) is linear:

- **approximation**: find coefficient $x \in \mathbf{R}^n$ so the linear combination

$$Ax = x_1 a_1 + \cdots + x_n a_n$$

  of the columns $a_1, \ldots, a_n \in \mathbf{R}^m$ of A is as close as possible to b

- **model fitting**: express residual in the rows $\tilde{a}_i^T$ of A:

$$r_i = \tilde{a}_i^T x - b_i, \quad i = 1, \ldots, m$$

  - data points $(\tilde{a}_i, b_i)$, $i = 1, \ldots, m$
  - *feature vector*: $\tilde{a}_i \in \mathbf{R}^n$ and label $b_i \in \mathbf{R}$
  - *model prediction*: $\tilde{a}_i^T x$ (for label b_i)

  find model parameter $x \in \mathbf{R}^n$ so the prediction error $r_i = \tilde{a}_i^T x - b_i$ is small for all data points

- **geometric interpretation**:

$$\begin{array}{ll} \text{minimize} & \|z - b\| \\ \text{subject to} & z \in \mathcal{R}(A) = \{Ax \mid x \in \mathbf{R}^n\} \end{array}$$

  find point in $\mathcal{R}(A)$ that is closest to b

- **estimation**: assume *linear measurement model* $b = Ax + r$ with r as (random) noise, find x minimizes the noise r so best explains b

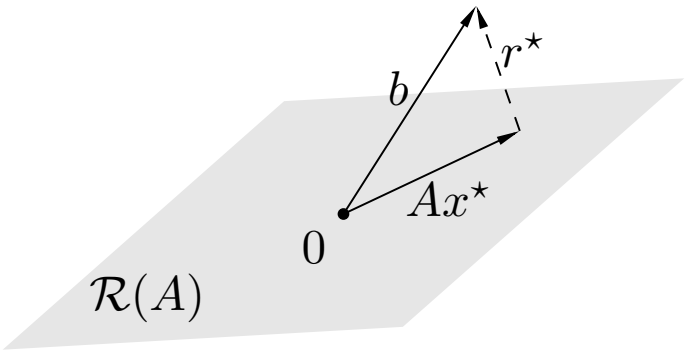
Least squares approximation

$$\text{minimize} \quad \|Ax - b\|_2^2 = r_1^2 + \cdots + r_m^2$$

- squared ℓ_2-*norm* (or *Euclidean norm*) of the residual
- rewrite objective as a convex quadratic function of x:

$$\|Ax - b\|_2^2 = x^T A^T A x - 2b^T A x + b^T b$$

- optimality condition: $\nabla \|Ax - b\|_2^2 = 2A^T A x - 2A^T b = 0$
- (unique) solution $x^\star = (A^T A)^{-1} A^T b$ (assume A tall, full rank)
- **geometric interpretation**:
  $x^\star$ is the projection of b onto $\mathcal{R}(A)$



- generalizes to nonlinear residual $f(x) - b$, but may not be convex

Least absolute residuals approximation

$$\text{minimize} \quad \|f(x) - b\|_1 = |r_1| + |r_2| + \cdots + |r_m|$$

- $\|u\|_1 = |u_1| + \cdots + |u_m|$ is the ℓ_1-*norm* of $u \in \mathbf{R}^m$
- convex optimization problem if f convex
- for linear $f(x) = Ax$ ($A \in \mathbf{R}^{m \times n}$),

$$\text{minimize} \quad \|Ax - b\|_1 = \sum_{i=1}^m |a_i^T x - b_i|$$

(a_i^T are rows of A) can be formulated as an LP:

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T t \\ \text{subject to} & -t \preceq Ax - b \preceq t \end{array}$$

(with variables $x \in \mathbf{R}^n$, $t \in \mathbf{R}^m$)

Least maximum residuals approximation

$$\text{minimize} \quad \|f(x) - b\|_\infty = \max\{|r_1|, \ldots, |r_m|\},$$

- $\|\cdot\|_\infty$: ℓ_∞-*norm* (or *Chebyshev norm*), maximum absolute value of the components
- also called *minimax* or *Chebyshev approximation*; convex if f convex
- for linear $f(x) = Ax$ ($A \in \mathbf{R}^{m\times n}$),

$$\text{minimize} \quad \|Ax - b\|_\infty = \max_{i=1,\ldots,m} |a_i^T x - b_i|$$

(a_i^T are rows of A) can be formulated as an LP:

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & -t\mathbf{1} \preceq Ax - b \preceq t\mathbf{1}, \end{array}$$

(with variables $x \in \mathbf{R}^n$, $t \in \mathbf{R}$)

Penalty function approximation

$$\text{minimize} \quad \phi(r_1) + \cdots + \phi(r_m) = \sum_{i=1}^m \phi(f_i(x) - b_i)$$

- ϕ: $\mathbf{R} \to \mathbf{R}$ is a (scalar) *penalty function*
- sometimes written as

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m \phi(r_i) \\ \text{subject to} & r = f(x) - b \end{array}$$

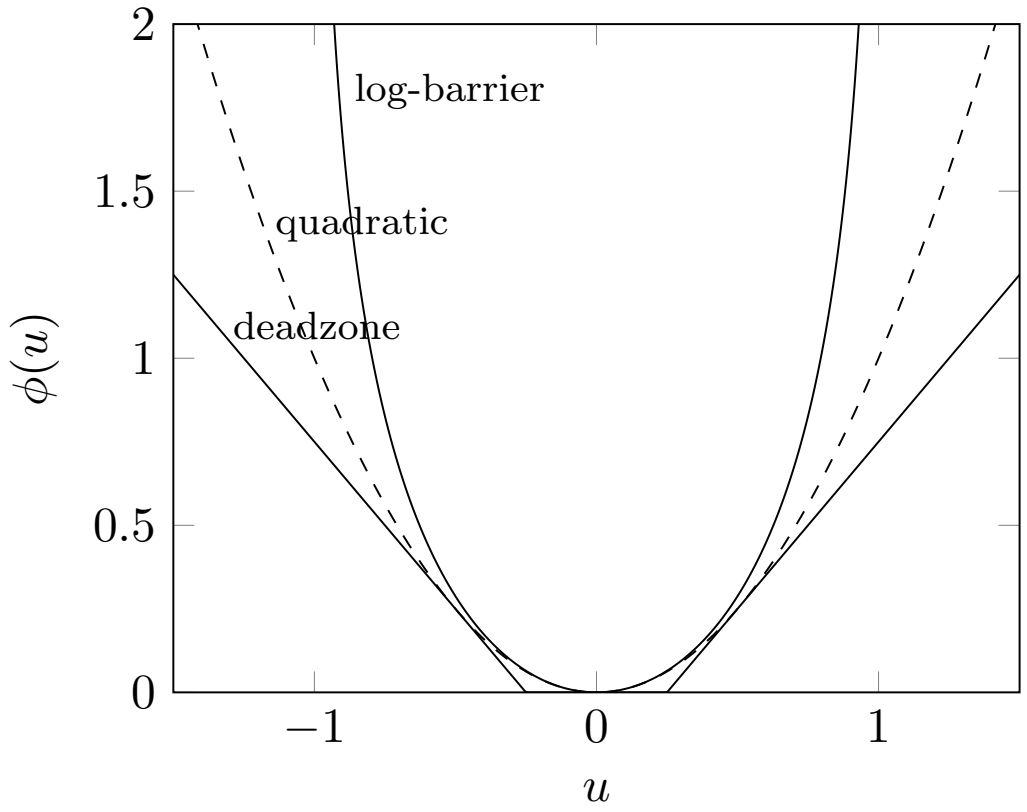
examples:

- ϕ associated with norms
  - ℓ_p-norm $\|r\|_p = \left(\sum_{i=1}^m |r_i|^p\right)^{1/p}$ $(p \in [1, \infty)) \iff \phi(r_i) = |r_i|^p$
    ($p = 1$: *absolute penalty*; $p = 2$: *quadratic penalty*)
  - ℓ_∞-norm $\|r\|_\infty \iff \phi(r_i) = \begin{cases} |r_i|, & |r_i| = \max\{|r_1|, \ldots, |r_m|\} \\ 0, & \text{otherwise.} \end{cases}$

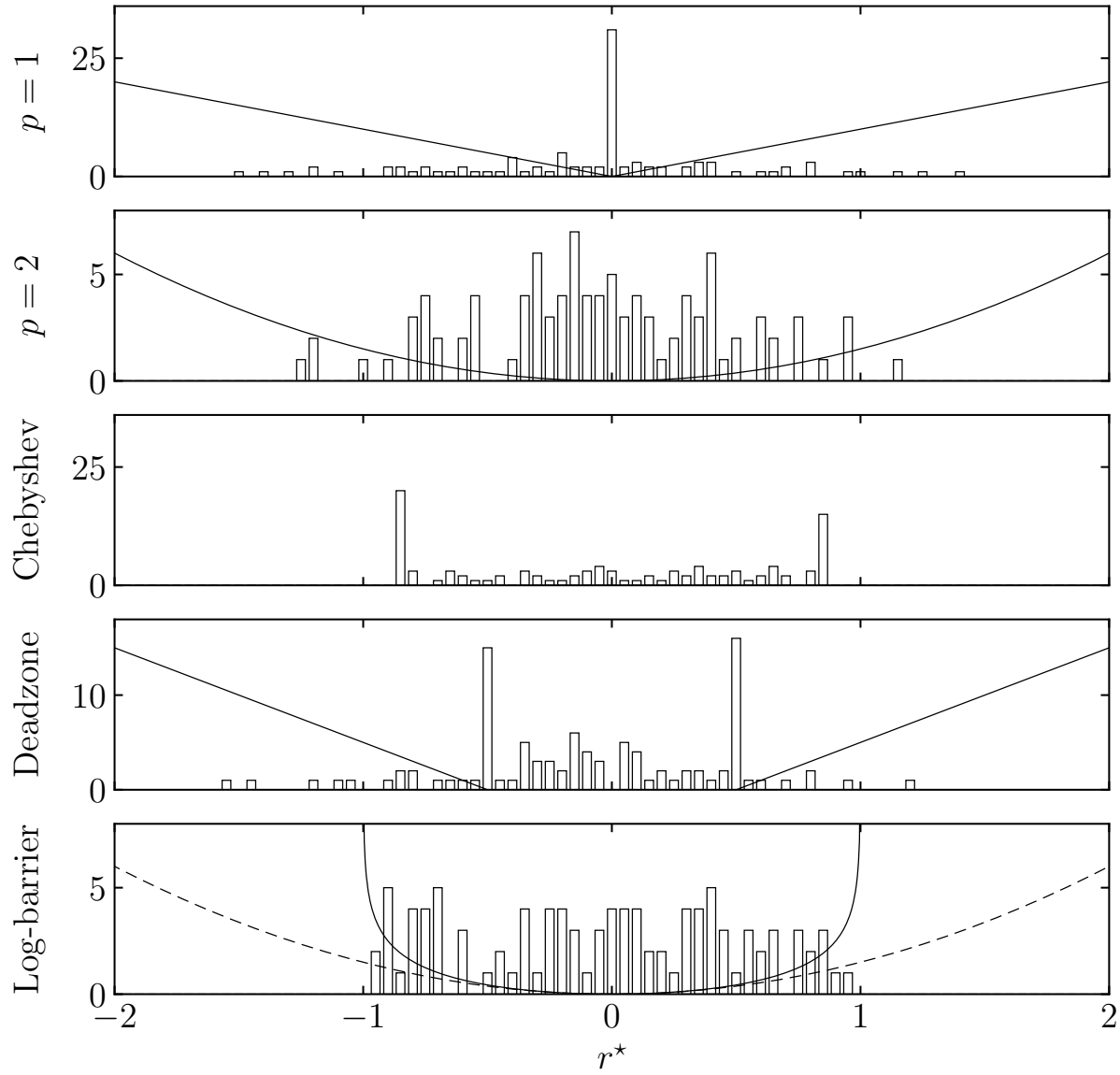
    (sometimes called *Chebyshev penalty*)

- *deadzone-linear penalty*: $\phi(u) = \begin{cases} |u| - \alpha, & |u| > \alpha \\ 0, & \text{otherwise} \end{cases}$

- *log-barrier penalty*: $\phi(u) = \begin{cases} -\alpha^2 \log(1 - (u/\alpha)^2), & |u| < \alpha \\ \infty, & \text{otherwise} \end{cases}$



numerical example: linear approximation of 100 data points in $\mathbf{R}^{30}$

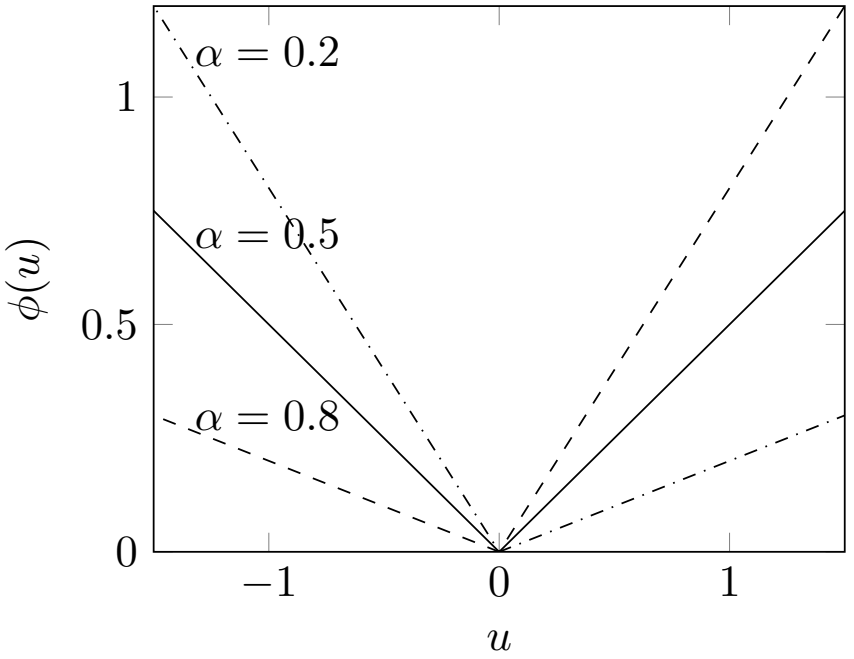


Quantile regression

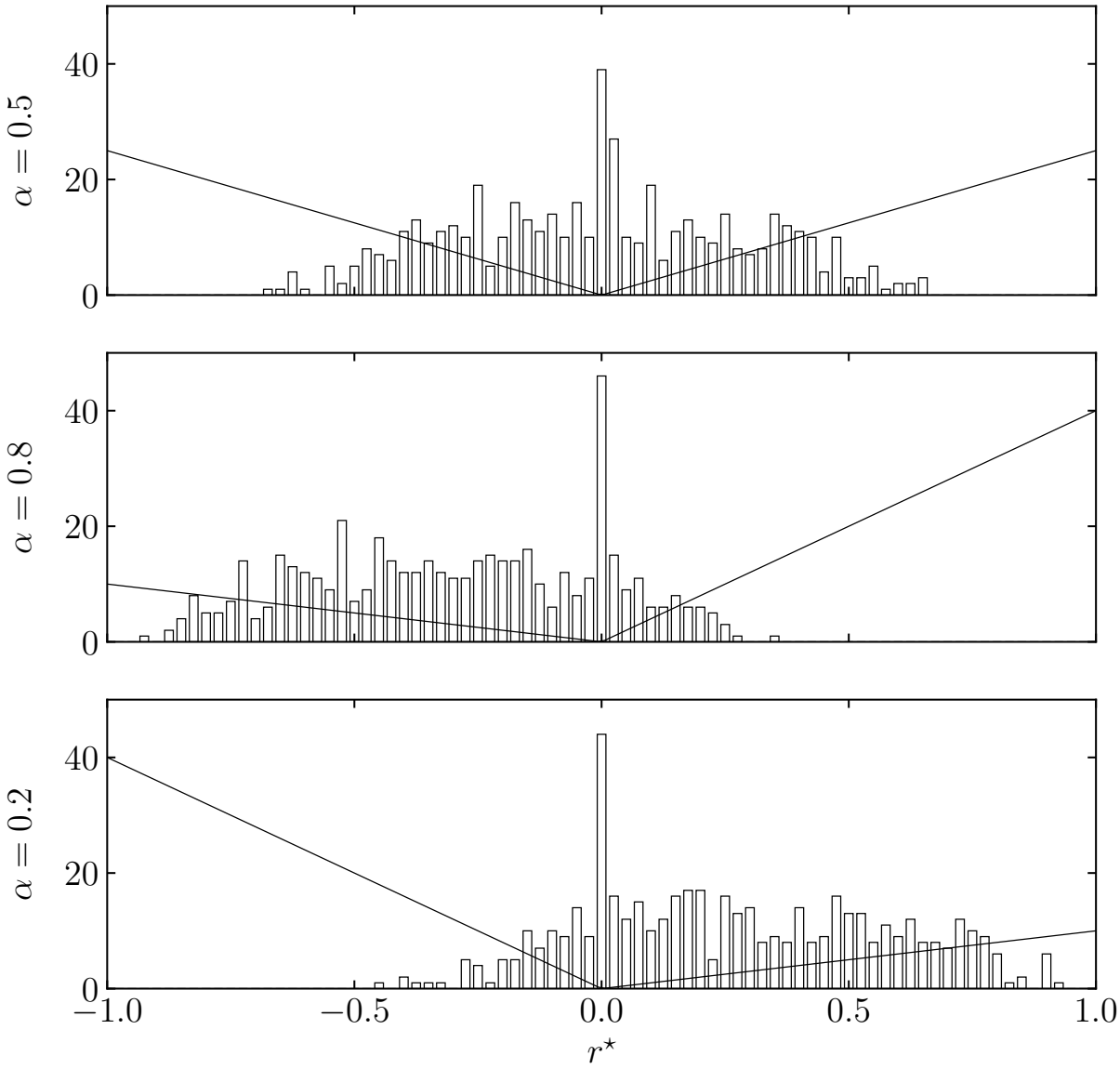
quantile penalty function:

$$\begin{aligned}\phi(u) &= \max\{\alpha u, (\alpha-1)u\} \\ &= \begin{cases} \alpha|u|, & u \geq 0 \\ (1-\alpha)|u|, & \text{otherwise} \end{cases}\end{aligned}$$

- parameter $\alpha \in (0,1)$
- $\alpha = 0.5$: absolute value penalty; otherwise tilted
- associated approximation problem called **quantile regression**



example: quantile regression of 500 data points in $\mathbf{R}^{30}$

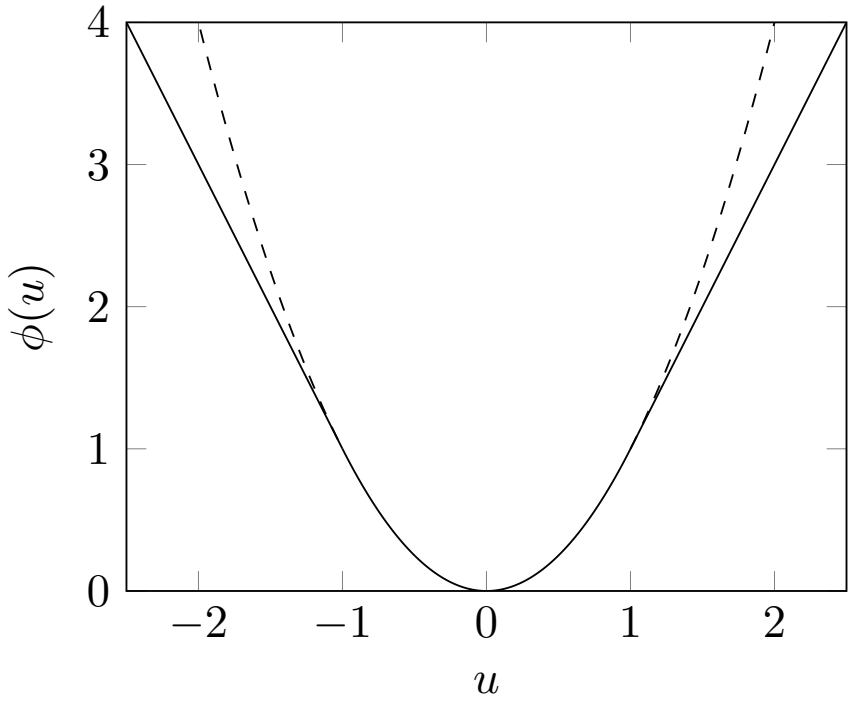


Robust regression

Huber penalty function (or *robust least squares penalty*):

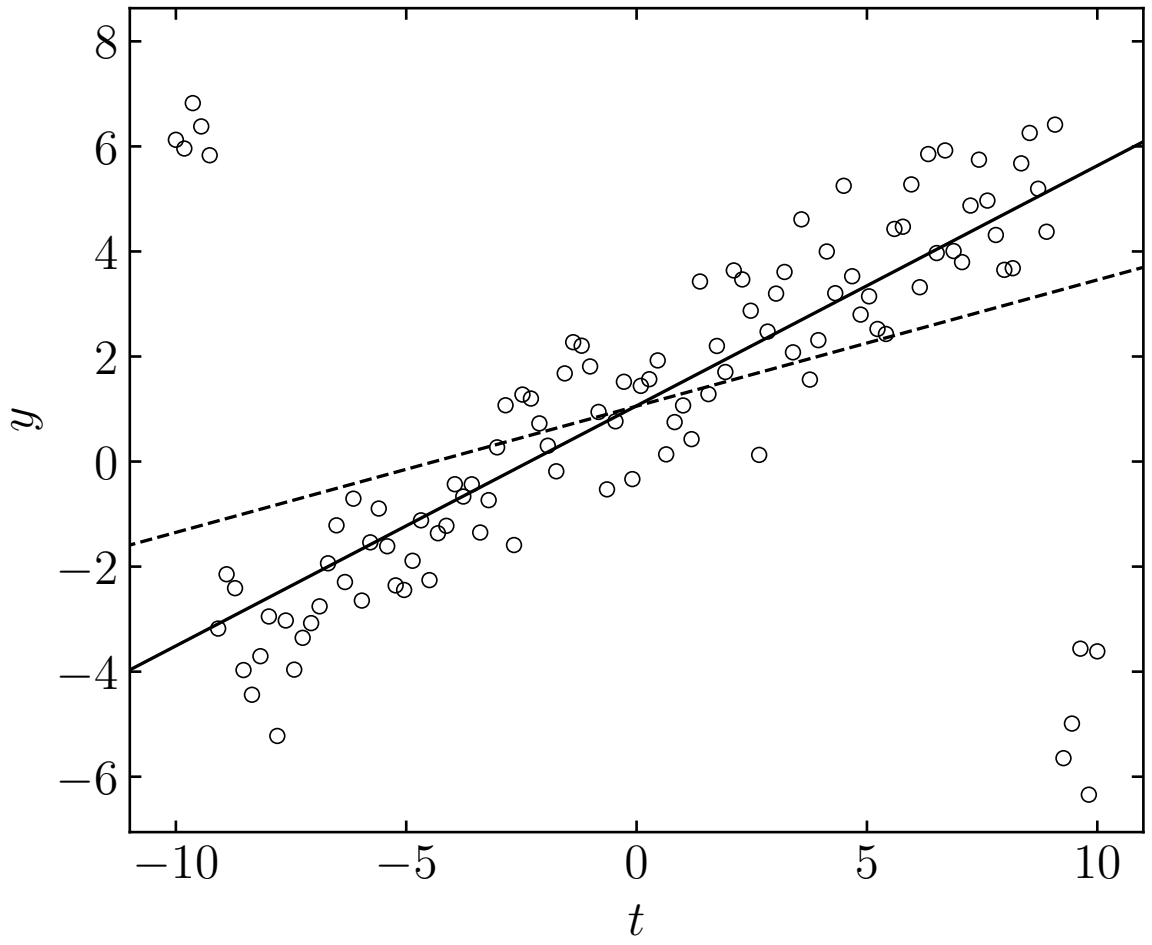
$$\phi(u) = \begin{cases} u^2, & |u| \leq M \\ M(2|u| - M), & \text{otherwise} \end{cases}$$

- parameter $M > 0$
- small residuals: quadratic penalty; large residuals: linear penalty
- hence not sensitive to outliers
- associated approximation problem called **robust regression**



example: robust regression in $\mathbf{R}^2$

- 100 points linearly distributed with random noise; 10 outliers
- dashed line shows ordinary least squares solution



Maximum likelihood estimation (MLE)

$$\text{maximize} \quad l(x) = \log p_x(y)$$

- variable $x \in \mathbf{R}^n$; data $y \in \mathbf{R}^m$ is given
- $p_x \colon \mathbf{R}^m \to \mathbf{R}_+$: density function of some random variable $y \in \mathbf{R}^m$ parameterized by $x \in \mathbf{R}^n$ (sometimes called *likelihood function*)
- $l \colon \mathbf{R}^n \to \mathbf{R}$ is the associated *log-likelihood function*

- find parameter x that maximizes the likelihood of observing the data y
- equivalent to minimizing the *negative log-likelihood* $-l(x)$
- convex program if l concave, $i.e.$, p_x log-concave in x

Linear approximation as MLE

linear measurement model: given dataset (a_i, y_i), $i = 1, \ldots, m$, assume

$$y_i = a_i^T x + v_i, \quad i = 1, \ldots, m$$

- $a_i \in \mathbf{R}^n$ are known *feature vectors*, $y_i \in \mathbf{R}$ are *responses*
- $x \in \mathbf{R}^n$ is unknown parameter, $v_i \in \mathbf{R}$ is random noise

assuming v_i are *independent and identically distributed* ($i.e.$, IID):

$$p_x(y) = \prod_{i=1}^m p(y_i - a_i^T x), \quad \text{so } l(x) = \sum_{i=1}^m \log p(y_i - a_i^T x)$$

- p: $\mathbf{R} \to \mathbf{R}_+$ is the density function of the noise v_i
- multiple possible choices of p leads to different MLE problems
- convex if p log-concave in x

Gaussian noise: $v_i \sim \mathcal{N}(0, \sigma^2)$ (with known $\sigma^2 > 0$), *i.e.*,

$$p(v) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{v^2}{2\sigma^2}\right)$$

let $A = \begin{bmatrix} a_1 & \cdots & a_m \end{bmatrix}^T \in \mathbf{R}^{m\times n}$,

$$\begin{aligned} l(x) &= \sum_{i=1}^{m} \log p(y_i - a_i^T x) \\ &= \sum_{i=1}^{m} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - a_i^T x)^2}{2\sigma^2}\right) \\ &= -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|Ax - y\|_2^2. \end{aligned}$$

so the MLE problem is equivalent to the least squares approximation

$$\text{maximize} \quad -\|Ax - y\|_2^2$$

Laplace noise: IID v_i following Laplace distribution (zero mean, known parameter $b > 0$), *i.e.*,

$$p(v) = \frac{1}{2b} \exp\left(-\frac{|v|}{b}\right)$$

log-likelihood function given by:

$$\begin{aligned} l(x) &= \sum_{i=1}^{m} \log p(y_i - a_i^T x) \\ &= \sum_{i=1}^{m} \left(-\log(2b) - \frac{|y_i - a_i^T x|}{b}\right) \\ &= -m \log(2b) - \frac{1}{b}\|Ax - y\|_1 \end{aligned}$$

MLE problem is equivalent to the least absolute residuals approximation

$$\text{maximize} \quad -\|Ax - y\|_1$$

uniform noise: IID $v_i \sim \mathcal{U}(-b, b)$ (with known $b > 0$), *i.e.*,

$$p(v) = \begin{cases} 1/(2b), & |v| \leq b \\ 0, & \text{otherwise} \end{cases}$$

log-likelihood function given by:

$$l(x) = \sum_{i=1}^{m} \log p(y_i - a_i^T x) = \begin{cases} -m \log(2b), & \|Ax - y\|_\infty \leq b \\ -\infty, & \text{otherwise} \end{cases}$$

any x satisfying $\|Ax - y\|_\infty \leq b$ is an MLE solution

in the most general case, any penalty function approximation problem

$$\text{minimize} \quad \textstyle\sum_{i=1}^{m} \phi(y_i - a_i^T x)$$

can be interpreted as MLE with IID noise having density function

$$p(v) = \frac{\exp(-\phi(v))}{\int \exp(-\phi(z))\ dz}$$

Probabilistic classification

logistic model: given feature $a_i \in \mathbf{R}^n$ with labels $y_i \in \{0, 1\}$, assume

$$\begin{array}{l} \mathbf{prob}(y_i = 1) = p_i \\ \mathbf{prob}(y_i = 0) = 1 - p_i \end{array}, \qquad p_i = \frac{\exp(a_i^T x)}{1 + \exp(a_i^T x)} \in (0, 1)$$

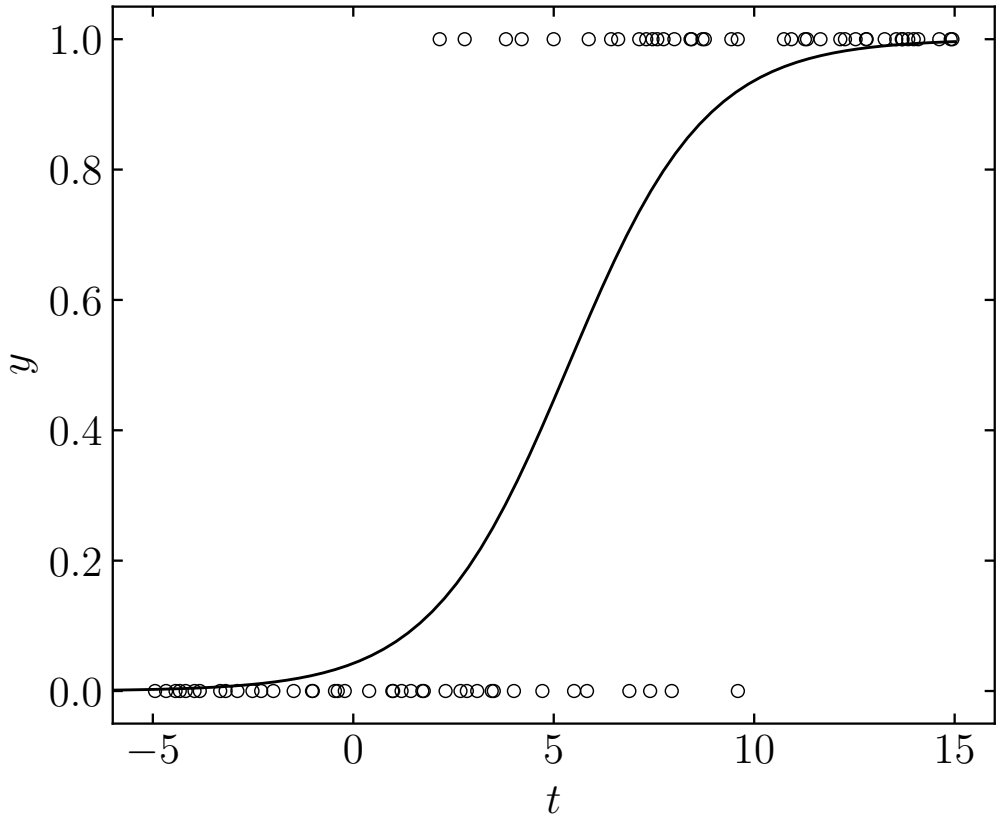
- estimate unknown parameter $x \in \mathbf{R}^n$

log-likelihood: assume $y_1, \ldots, y_k = 1$ and $y_{k+1}, \ldots, y_m = 0$,

$$\begin{aligned} l(x) &= \sum_{i=1}^{k} \log p_i + \sum_{i=k+1}^{m} \log(1 - p_i) \\ &= \sum_{i=1}^{k} \log \left(\frac{\exp(a_i^T x)}{1 + \exp(a_i^T x)} \right) + \sum_{i=k+1}^{m} \log \left(\frac{1}{1 + \exp(a_i^T x)} \right) \\ &= \sum_{i=1}^{m} \Big(y_i a_i^T x - \log(1 + \exp(a_i^T x)) \Big) \end{aligned}$$

- maximize $l(x)$ is a convex optimization problem

example: 80 data points in $\mathbf{R}^2$ with $a_i = (t_i, 1)$



solid curve shows $\mathbf{prob}(y = 1)$ for MLE solution $x^\star$

Counting problems

Poisson model: given feature $a_i \in \mathbf{R}^n$ with count label $y_i \in \mathbf{Z}_+$, assume

$$\mathbf{prob}(y_i = k) = \frac{\lambda_i^k \exp(-\lambda_i)}{k!}, \quad \log \lambda_i = a_i^T x$$

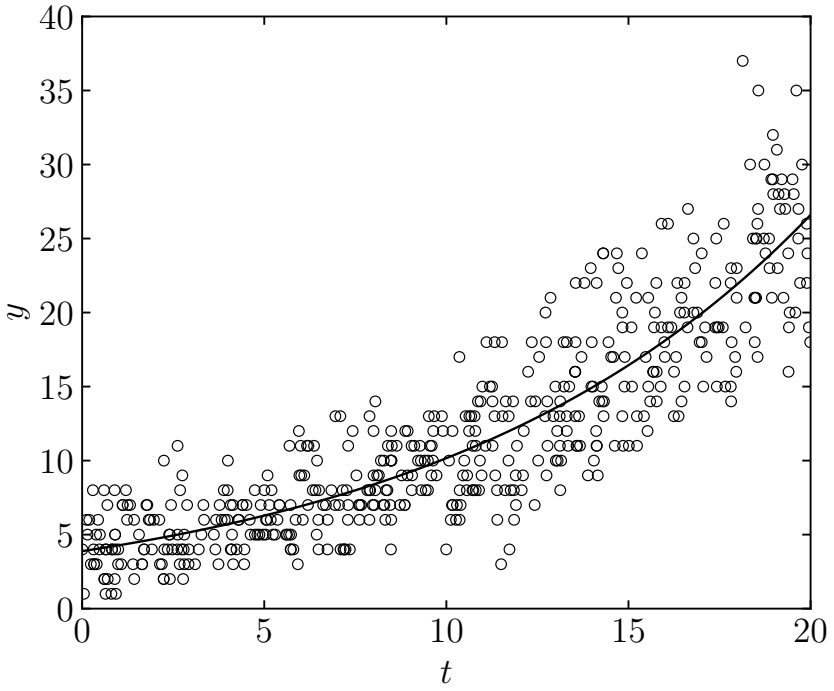
log-likelihood of parameter $x \in \mathbf{R}^n$:

$$l(x) = \log \prod_{i=1}^m \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}$$

$$= \sum_{i=1}^m (y_i \log \lambda_i - \lambda_i - \log(y_i!))$$

$$= \sum_{i=1}^m \left(y_i a_i^T x - \exp(a_i^T x) - \log(y_i!) \right)$$

- maximize $l(x)$ is convex



Gaussian covariance estimation

given Gaussian IID samples $y_1, \ldots, y_m \in \mathbf{R}^n$ for random $y \in \mathbf{R}^n$ with mean zero and covariance $X = \mathbf{E}\, yy^T \succ 0$, log-likelihood of X is

$$\begin{aligned} l(X) &= \sum_{i=1}^m \log p_X(y_i) = \sum_{i=1}^m \left(-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det X - \frac{1}{2} y_i^T X^{-1} y_i \right) \\ &= -\frac{mn}{2}\log(2\pi) - \frac{m}{2}\log\det X - \frac{1}{2}\mathbf{tr}\left(X^{-1} \sum_{i=1}^m y_i y_i^T \right) \end{aligned}$$

- $p_X : \mathbf{R}^n \to \mathbf{R}_+$ is Gaussian density function with covariance X

$l(X)$ nonconcave, but let $S = X^{-1}$ gives convex MLE problem:

$$\begin{array}{ll} \text{maximize} & \log\det S - \mathbf{tr}(SY) \\ \text{subject to} & S \succ 0 \end{array}$$

- $Y = (1/m)\sum_{i=1}^m y_i y_i^T \succeq 0$ is the sample covariance matrix
- if $Y \succ 0$, unique solution $S^\star = Y^{-1}$, otherwise unbounded above

Nonparametric distribution estimation

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \succeq 0, \quad \mathbf{1}^T x = 1 \end{array}$$

- variable $x \in \mathbf{R}^n$ represents a discrete distribution over n outcomes
- convex optimization problem if f convex

maximum entropy estimation:

$$f(x) = \sum_{i=1}^{n} x_i \log x_i$$

- solution $x^\star = (1/n)\mathbf{1}$ is the uniform distribution over n outcomes

minimum KL-divergence estimation:

$$f(x) = D_{\mathrm{kl}}(x, q) = \sum_{i=1}^{n} x_i \log(x_i/q_i)$$

- $q \in \mathbf{R}^n$ ($q \succeq 0$, $\mathbf{1}^T q = 1$) is a given reference distribution
- solution $x^\star = q$

maximum likelihood estimation given IID samples

$$z_i \sim x, \quad z_i \in \{1, \ldots, n\}, \quad i = 1, \ldots, m$$

from unknown distribution $x \in \mathbf{R}^n$:

$$\begin{array}{ll} \text{minimize} & -l(x) = -\sum_{i=1}^m \log x_{z_i} \\ \text{subject to} & x \succeq 0, \quad \mathbf{1}^T x = 1 \end{array}$$

- solution $x_i^\star = c_i/m$, where $c \in \mathbf{R}^n$ is the sample count vector

bounding problems:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n g(c_i) x_i \\ \text{subject to} & x \succeq 0, \quad \mathbf{1}^T x = 1 \end{array}$$

- find a lower bound on expected value of g: $\mathbf{R} \to \mathbf{R}$ over $\{c_1, \ldots, c_n\}$ under distribution x
- similar for lower bounding, *e.g.*, probability of events, sum of k largest probabilities, . . .

Discrimination

given $\{x_1, \ldots, x_M\}, \{y_1, \ldots, y_N\} \subseteq \mathbf{R}^n$, find $f : \mathbf{R}^n \to \mathbf{R}$ such that

$$f(x_i) > 0 \text{ for all } i = 1, \ldots, M, \qquad f(y_i) < 0 \text{ for all } i = 1, \ldots, N$$

i.e., the 0-sublevel set of f separates the two sets

linear discrimination:

$$\begin{array}{ll} \text{find} & (a, b) \\ \text{subject to} & a^T x_i - b > 0, \quad i = 1, \ldots, M \\ & a^T y_i - b < 0, \quad i = 1, \ldots, N \end{array}$$

- constraints homogeneous in (a, b), so equivalent to

$$\begin{array}{ll} \text{find} & (a, b) \\ \text{subject to} & a^T x_i - b \geq 1, \quad i = 1, \ldots, M \\ & a^T y_i - b \leq -1, \quad i = 1, \ldots, N \end{array}$$

quadratic discrimination:

$$\begin{array}{ll} \text{find} & (P, q, r) \\ \text{subject to} & x_i^T P x_i + q^T x_i + r \geq 1, \quad i = 1, \ldots, M \\ & y_i^T P y_i + q^T y_i + r \leq -1, \quad i = 1, \ldots, N \end{array}$$

- LP of variables $P \in \mathbf{S}^n$, $q \in \mathbf{R}^n$, $r \in \mathbf{R}$
- $P \succ 0$ gives an ellipsoidal separator

